世界互联网大会
World Internet
Conference

# 发展负责任的生成式人工智能共识

# CONSENSUS ON DEVELOPING RESPONSIBLE GENERATIVE ARTIFICIAL INTELLIGENCE

世界互联网大会人工智能工作组
2023年11月

World Internet Conference Working Group on Artificial Intelligence
November 2023

# 发展负责任的生成式人工智能共识
# Consensus on Developing Responsible Generative Artificial Intelligence

## I.总则
## General

01 **发展负责任的生成式人工智能应始终致力于增进人类福祉，坚持以人为本，推动人类经济、社会和生态可持续发展。**应正确认识生成式人工智能所蕴含的巨大潜力和可能风险，遵循统筹发展和安全、平衡创新与伦理、均衡效益与风险的理念，推动生成式人工智能负责任的发展。一方面，应积极推动创新、可持续、包容开放的发展，提升生成式人工智能算力高效、数据高质、算法创新、人才多元、生态开放的能力；另一方面，以高度负责任的态度发展可靠可控、透明可释、数据保护、多元包容、明确责任、价值对齐的生成式人工智能。

01 **Developing responsible Generative Artificial Intelligence(GenAI) should consistently commit to enhancing human well-being, adhering to a people-centered approach, and promoting the sustainable development of the economy, society, and ecological environment**. It is important to accurately recognize the immense potential and possible risks of GenAI. By adhering to a holistic approach that harmonizes development with safety and security, balances innovation with ethical considerations, and fairly evaluates benefits and risks, we can promote the responsible development of GenAI. On the one hand, innovation-driven, sustainable, inclusive, and open development of GenAI should be promoted, and the efficient computing power, high-quality data, innovative algorithm, diverse talents, and an open ecosystem associated with GenAI should be enhanced. On the other hand, GenAI characterized by reliability and controllability, transparency and explainability, data protection, diversity and inclusiveness, accountability, and value alignment should be developed with a highly responsible attitude.

## II.促进生成式人工智能发展
## Promote the Development of GenAI

02 **积极倡导并稳妥推进生成式人工智能的可持续发展。一是保证经济可持续性。**应确保生成式人工智能提高生产力和创造就业机会，提高资源使用效率，实现数实融合的循环经济，推动科技创新，促使经济结构向更高附加值的转变；**二是保证社会可持续性。**应确保生成式人工智能公平与平等的使用，实现全社会对其的共享共治；**三是保证环境的可持续性。**生成式人工智能的发展应实现对于自然资源的可持续管理和使用，鼓励采用绿色能源驱动基础设施、提高能源转化效率、绿色开发算法模型应用，降低温室气体排放，实现绿色发展。

02 **Actively advocate and prudently promote the sustainable development of GenAI. Firstly, ensure economic sustainability**. It is vital to ensure that GenAI can enhance productivity and create employment opportunities, improve the efficiency of resource utilization, achieve a circular economy that integrates digital technologies with the real economy, propel scientific and technological innovation, and facilitate the transformation to an economic structure with higher value added. **Secondly, ensure social sustainability**. It is crucial to ensure the fair and equitable use of GenAI and to promote co-construction and co-governance at the societal level. **Thirdly, ensure environmental sustainability**. The development of GenAI should promote the sustainable management and use of natural resources, encourage the use of green energy-driven infrastructure, and enhance green R&D of models and applications, to reduce greenhouse gas emissions and achieve green development.

03 **构建有益于生成式人工智能健康有序发展的良好环境。一是建立和完善相关的伦理原则和法律法规，**重点审视知识产权法律制度，探索人工智能生成物的权利归属方案，对其进行恰当的管理和保护。**二是构建包容、扶持、前瞻、可预期的政策环境。**为前沿应用孵化构建一个包容的创新环境，为规模推广营造一个优良的营商环境，为赋能经济社会发展搭建一个稳健的监管环境。**三是加强国际交流与合作。**生成式人工智能的发展需要全球各利益相关方秉持共商共建共享理念，以开放协作态度和举措，开展跨国家、跨领域、跨文化交流与协作，推动形成具有广泛共识的国际评测及标准体系，确保各国共享生成式人工智能的技术惠益。

03 **Create an environment that is conducive to healthy and orderly development of GenAI. Firstly, establish and optimize relevant ethical principles, laws, and regulations**. Legal assessment of the adequacy of the current intellectual property laws should be done with concerned stakeholders to ensure proper attribution of rights to AI-generated objects and to appropriately manage and protect them. **Secondly, establish an inclusive, supportive, forward-looking, and predictable policy environment**. GenAI requires an inclusive innovation environment for incubating its cutting-edge applications, an excellent business environment for deploying them at scale, and a robust regulatory environment to drive economic and social development. **Thirdly, strengthen international communication and cooperation**. The development of GenAI requires all global stakeholders to uphold the principle of consultation, contribution, and shared benefits, to engage in transnational, cross-disciplinary, and cross-cultural communication and cooperation with open and collaborative attitudes and actions, to jointly establish an international assessment and standard system with broad consensus, to ensure that all countries can share in the technological benefits of GenAI.

04 **提升生成式人工智能研发及规模应用的能力。一是构建开放共享、普惠包容的算力资源。**应推动算力的合理分配与高效利用，降低科技创新的门槛，确保不同地区、不同规模的企业及个人都能获得必要的计算资源。**二是推动负责任的数据共享。**应鼓励推广高质量数据的共享流动，增强公共数据资源供给，保障数据安全共享与合规利用，提升各领域数据治理水平。**三是完善算法创新的设施条件。**应前瞻谋划、统筹布局各类平台和开放共享服务网络建设，鼓励算法和基础模型在安全的基础上开源开放，加强跨行业、跨领域协作，推动产学研结合，形成算法创新的良性生态。**四是全面加强人才能力建设。**针对从业者，应建立人才交流平台，促进互学互鉴与知识共享，设计并实施涵盖多层次、多领域的教育培训项目，增进不同领域技术供需双方的交流与学习。针对公众，应加强科普、教育及培训，提供准确认知，提升数字素养，促进生成式人工智能的普遍接入。**五是推动重点领域应用赋能。**推动生成式人工智能与各行业数字化场景深度融合，实现应用迭代创新，促进生成式人工智能技术成果在重点领域的应用赋能。

04 **Enhance the capabilities of R&D and scaling the application of GenAI. Firstly, build open, shared, and inclusively computing resources that are beneficial to all**. Efforts should be made to promote the rational allocation and efficient use of computing power, reducing the barriers to scientific and technological innovation. This will enable enterprises of different regions and tiers to access the computational resources they need. **Secondly, promote responsible data sharing**. Efforts should be made to encourage the sharing and flow of high-quality data, increase the supply of public data resources, ensure the safe sharing and compliant use of data, and strengthen the level of data governance in all fields. **Thirdly, optimize the facility system for algorithm innovation**. Forward-looking planning and overall arrangement for the construction of various platforms and open shared service networks should be enacted, secure open-sourcing of algorithms and basic models should be encouraged, cross-industry and cross-domain collaboration should be strengthened, and the integration of industry, academia, and research should be promoted, to form a virtuous ecosystem for algorithmic innovation. **Fourthly, comprehensively strengthen talent capacity building**. For practitioners, a talent exchange platform should be established to promote mutual learning and knowledge sharing, and multi-level,

multi-domain education and training programs should be designed and implemented to enhance communication and learning between technology providers and users in different fields. For the public, science popularization, education, and training should be enhanced to provide accurate knowledge, improve the digital literacy, and promote universal access to GenAI. **Fifthly, enhance the empowerment of critical application areas**. Promote the seamless integration of GenAI with digital scenarios across various industries, enabling application innovation and iteration, and driving the adoption of GenAI in key sectors.

## III.提升生成式人工智能的负责任治理能力
## Enhance the Capacity for Responsible Governance of GenAI

05 **发展安全可靠的生成式人工智能，确保全生命周期内可控地运行。一是提升安全稳健性和生成准确性。**增强生成式模型防御提示攻击、注入攻击等能力，不断提高稳健性和抗干扰能力。探索内容生成可控的技术或解决方案，确保生成的信息内容尽可能准确。**二是确保人类知情与控制。**确保人类知悉其在与生成式人工智能交互，确保生成式人工智能系统可被人类监督和及时接管。**三是避免技术滥用与恶意使用。**避免用户过度依赖生成式人工智能，减轻其对人类创新力与主体性的负面影响。避免故意或非故意地使用生成式人工智能伤害社会与公众利益。

05 **Develop safe and reliable GenAI to ensure controlled operation throughout its lifecycle. Firstly, improve the safety, robustness, and generative accuracy of GenAI**. Enhance the capabilities of generative models to defend against prompted attacks, and other types of attacks, while continuously improving their robustness and anti-interference capabilities. Explore technologies or solutions for ensuring controlled content generation to ensure that the generated information is as accurate as possible. **Secondly, ensure humans are well-informed and in control**. It is important to ensure that humans are aware of their interactions with GenAI and that GenAI systems are supervised and controlled by humans in a timely manner. **Thirdly, avoid the abuse and misuse of technologies**. Mitigate excessive dependence on GenAI by users and reduce its negative impact on human innovation and subjectivity. Avoid intentionally or unintentionally using of GenAI to harm society and the public interest.

06 **增强生成式人工智能系统的透明度与可解释性，提升人类对其理解和信任。一是提升透明度，**鼓励在安全的基础上披露生成式人工智能系统的能力及局限性，以及决策过程及技术意图；建立外部监督与反馈渠道，并不断做出改进。**二是增强可解释性，**推动生成式人工智能的可解释性研究，探索自适应场景和风险水平的强可解释性技术路线，增进人类信任，提升应用接受度。

06 **Enhance the transparency and explainability of GenAI systems to improve human understanding and trust. Firstly, improve transparency**. Encourage disclosure of capabilities and limitations of GenAI systems, as well as decision-making processes and technical intentions, based on safety. Establish external supervision and feedback channels, and continuously make improvements. **Secondly, enhance explainability**. Promote research on the explainability of GenAI, explore robust explanatory technical pathways for adaptive scenarios and risk levels, to enhance human trust and improve application acceptance.

07 **强化生成式人工智能数据治理，加强数据安全，尊重和保护个人隐私。一是强化数据治理，**避免训练数据的非法收集、滥用和泄漏等问题，采取有效措施提高训练数据质量。**二是加强个人信息与隐私保护。**生成式人工智能训练数据涉及个人数据时应依法获得用户知情和同意，确保生成内容不侵犯个人隐私。**三是探索隐私保护技术，**在构造生成式人工智能系统时，探索使用隐私计算等技术，防范数据泄露及滥用风险。

**07 Strengthen GenAI data governance and data security, respecting and protecting individual privacy. Firstly, reinforce data governance**. Problems such as illegal collection, abuse, and breach of training data should be avoided, and effective measures should be taken to improve the quality of training data. **Secondly, enhance the protection of personal information and privacy**. When GenAI training data involves personal data, the owner should be informed and consulted for consent to ensure that the generated content does not infringe on individual privacy. **Thirdly, explore technologies for privacy protection**. When developing GenAI systems, it is important to explore the use of privacy-preserving computing and other privacy protection technologies to mitigate the risk of data leakage and misuse.

**08 确保生成式人工智能的开放包容和公平普惠。一是确保技术多元包容，**保障生成式人工智能的训练数据、应用场景具有必要的多元性，避免产生对特定群体或个人的偏见或歧视。**二是促进技术公平普惠，**降低生成式人工智能的成本和使用门槛，提升其可得性和易用性，推动人类社会共享生成式人工智能带来的益处，促进社会公平和机会均等，弥合数字鸿沟。

**08 Ensure the openness, inclusiveness, and fairness of GenAI. Firstly, ensure diversity and inclusiveness in technology**. GenAI training data and application scenarios should be diverse to avoid bias and discrimination against specific groups or individuals. **Secondly, promote a fair and equitable access to the technology**, reduce the cost and usage barriers of GenAI, enhance its accessibility and usability, advocate for the sharing of benefits that GenAI brings to human society, promote social fairness and equal opportunities, and bridge the digital divide.

**09 明确生成式人工智能的归责体系，增强系统可追溯性。一是明晰归责体系，**科学设计不同类型主体在生成式人工智能设计、训练、优化、部署、应用等全生命周期的权利义务和归责体系，确保在损害发生时可问责；**二是构建追溯机制，**鼓励成立并完善人工智能伦理委员会，确保决策过程及结果可追溯；**三是探索治理沙盒等创新友好型治理工具体系，**为生成式人工智能提供试错空间，支持负责任的创新探索。

**09 Clarify the mechanisms for accountability for GenAI and enhance system traceability. Firstly, clarify the mechanisms for accountability**, including the appropriate right obligations and accountability mechanisms for various categories of scientific design subjects throughout the life cycle of GenAI. This includes design, training, optimization, deployment, application, and so forth. These mechanisms should ensure liability in the event of any damage occurring. **Secondly, establish a traceability system**. AI ethics committees should be encouraged to establish and improve. Ensure GenAI results are traceable. **Thirdly, explore innovation-friendly governance tools like sandbox**, to provide experimental space for GenAI, fostering responsible innovative exploration.

**10 推动生成式人工智能更好地理解人类意图、遵循人类指令并符合人类的伦理道德。一是探索价值对齐研究，**加强生成式人工智能价值对齐理论探索、技术研究和工具研发，提升人类设计、理解和监督生成式人工智能模型的能力；**二是提升价值对齐技术，**提升生成式人工智能的训练数据质量，采取人工或自动化检测、红队测试、水印标记、内容过滤等手段，增强其与人类价值的一致性。

**10 Promote GenAI to better comprehend human intentions, follow human directives, and align with human ethics. Firstly, explore research on value alignment**. Strengthen GenAI value alignment theory exploration, technological research, and R&D of tools, and improve the ability of humans to design, understand and supervise GenAI models. **Secondly, enhance value alignment technologies**. Improve the quality of GenAI' s training data, adopting manual or automated detection, red team testing, watermarking, content filtering, and other methods to improve its consistency with human values.

世界互联网大会
World Internet
Conference

工作组联系方式/Contact us

research@wicinternet.org